COMPARATIVE ANALYSIS OF UNSTRUCTURED DATA USING FWS, FIMDO, WFUPAC ALGORITHM

K.V.Kanimozhi[1], Pradeep Kumar S[2] ,R.Beaulah Jeyavathana[3]
[1,3]Assistant Professor (SG), Department of Computer Science and Engineering,
Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences,
Chennai, India.
[2]Associate Professor, Department of Computer Science and Engineering,
Rajalakshmi Institute of Technology, Chennai, India.
email- kani.kalai4@gmail.com

Abstract. In traditional unstructured text clustering techniques mostly they use vector space model usually considers all the text documents as bags of words where the word sequence are not considered for efficient clustering. For Cluster quality the order of terms in the document collection plays a main role in which vector space model do not support. Hence recent days for the text clustering usually done through frequent item based. This paper analysis the different techniques like Frequent word sequence(FWS),Frequent item based on maximum document occurrence(FIMDO) and weighted frequent utility pattern agglomerative clustering(WFUPAC) and evaluates the input datasets like newsgroup and Reuters dataset with varying size. The result proves that the weighted frequent utility pattern agglomerative clustering(WFUPAC) outperforms when compared to Frequent word sequence (FWS) and Frequent item based on maximum document occurrence(FIMDO).Thus enhances the accuracy of text clustering in big data environment.

Introduction

Clustering analysis is commonly known for data discretization, most of the textual documents not only contains keywords it also consists of paragraph, sentences, error reports, product specification, notes, summary reports, research articles, newsgroups etc. The importance of clustering text group similar documents efficiently by maximizing intra class similarity and minimize interclass similarity. Clustering mainly used to find the valid and consistent organization of data and to uncover natural groupings. Normally the unstructured text data are incomplete, redundant, noisy and inconsistent. Thus pre-processing can be done for the data which helps to improve the quality of clustering results. The different pre-processing steps performed in text mining are tokenization helps to split the long sentence lines into individual terms. Normally the text clustering algorithms are classified into two general types namely, portioning technique and hierarchal technique. K-Means algorithm is an example of partitioning technique where it uses heuristic function. Initial centroids are chosen then uses distance function to calculate the similarity measures and until k-optimum clusters are obtained the above steps are repeated which uses only one level partitioning. In hierarchal method of clustering here it takes every text document as a cluster. Cluster overlapping is restricted in this method. It mainly uses various types of distance formula to measure the similarity between textual documents and finally combines the closest pair. These merging steps are repeated until the specified number of clusters is obtained. Then the Bisecting k-means algorithm is used which merges the advantages of both k-means and hierarchal clustering techniques and shows the improved accuracy and efficiency in cluster formation. These three conventional clustering methods has many disadvantages like, • These techniques does not reduce huge dimension as it leads to huge number of individual terms during tokenization step and processing the clusters. • These techniques works well for formatted input databases and importance not given to main characteristics of unstructured text documents databases. • It is hard to suggest the required number of clusters before the clustering process. • In order to make use of cluster output efficiently, every cluster must contain detail description of content. • It takes huge computation time for repeating the algorithm steps until to obtain k-number of clusters. Hence most of the recent research focuses on frequent item based clustering instead of using distance measure function. In the remaining section literature work based on existing techniques shown in section 2 using frequent item based clustering on text data. Section 3 covers about problem statement and proposed methodologies where it discusses about three different techniques and its details and section 4 consists of experiments done on the three different techniques and results are compared. Section 5 refers to the conclusion.

2. Related Work

The paper discusses [4,13] about various frequent clustering algorithm are calculated by verifying with some of the percentage provided in text documents. The comparison is performed between above two algorithms with bisecting k-means and proved that better accuracy in clustering the documents.

A new algorithm was implemented named as frequent pattern tree structure gives more details about frequent patterns using FP growth technique and proved that outperforms the Apriori technique to find frequent patterns. The paper uses the concept called hierarchal based clusters [1,2] and it uses a an specific type of distance calculation techniques. The frequent weighted utility item sets (FWUI) [5,6,7] discusses about the calculation of term vectors, inverse document frequency and the Modification Weighted Itemset Tidset (MWIT)-FWUI techniques used for matrix calculation and then maximum capturing methods are used to obtain the final clusters. The model is proposed for analyzing and clustering a geographical locations.[5,6,7].

Feature selection algorithm with weight scheme and dynamic dimension reduction is proposed for text clustering.[8] A neural feedback clustering method [9] combines with bidirectional long short-term memory and convolution neural network with k-means is proposed and analysis is done to achieve greater results in clustering.

Nowadays to inorder to icrease the efficiency novel algorithm with optimization techniques are used.[11], DBSTexC, a novel density- based clustering method, DBSCAN method [10], Spatial Indexing[10],Vector space model using k-means [21], multi-objectives-based method [12], particle swarm optimization algorithm using k-means[13], work for market prediction, review of various versions of KH methods [14], modified Multi-Swarm PSO (MSPSO) [15], various text clustering [14,15,16] are performed and analysis is done.

3 Problem statement and proposed solution

Frequent word sequence:

The text document is divided into individual words by tokenization. The user mention the minimum support count for the frequent item set so that the frequent words occurred in all documents are extracted as per min count. The words are represented as w1, w2, w3, w4… from documents d, In a particular order the continuation of words are mentioned where the desired number of percentage documents must support sequence. The algorithm steps initially starts with minimum of two word sets then for remaining all word sequence using data structure method called generalized suffix tree also called as (GST). The words which are not frequent are eliminated, only the words which support minimum support count is considered.

Once the compact documents obtained after finding the frequent word sequence, it helps to reduce huge dimension. Hence from frequent word sequence only compact documents obtained and remove other documents not in word sequence. Using the generalized suffix tree method, the tree is drawn only for compact documents containing frequent word sequence where a conventional method of suffix tree clustering tree is constructed for entire input documents. This method is very useful and more efficient compared to suffix tree clustering and other vector space model.

3.2. Finding frequent item based maximum document occurrence:

The algorithm is implemented in map reduce programming paradigm in hadoop where it calculates final clustering based on various steps like

*Performing the preprocessing by tokenization in order to obtain individual items and stops occurred are removed.

* Then the minimum count from every document is calculated, similarly repeated for all documents from every sentence using mapper and reducer task.

*Then the bivariate ngram techniques used to generate item sets from value 2 to n sets until minimum support count is reached.

*then the frequent item sets is calculated from those item sets.

*Next the document similarity matrix is determined using input matrix and outputs related documents values with as occurrences.

*Finally based on choosing the maximum occurred text documents centroid, remaining associated text documents re iteratively clustered and gives the output as final clusters.

3.3 Weighted Frequent Utility Pattern Agglomerative Clustering:

The algorithm is implemented [22, 23] using map and in different steps like

*Tokenization is used as initial step to obtain individual terms and stop words are removed.

*After obtaining set of terms ,we need to assign the weights to the terms then weight matrix are calculated.

* Then frequent weighted utility patterns mining are done.

* The Transaction Weighted Utility twu of a transaction tk is determined as,

$$\mathrm{twu}(\mathrm{t_k}) = \frac{\sum_{\mathrm{i,j \in S(t_k)}} \mathrm{w_j} * \mathrm{x_{k_{ij}}}}{|\mathrm{t_k}|}$$

* The weighted utility support wus of an itemset X is determined using equation

$$\mathrm{wus}(\mathrm{x}) = \frac{\sum_{\mathrm{t_k \in t(x)}} \mathrm{twu}(\mathrm{t_k})}{\sum_{\mathrm{t_k \in t}} \mathrm{twu}(\mathrm{t_k})}$$

Then Weighted Frequent Utility Pattern Agglomerative Clustering is performed and finally obtains the clusters.

4. Results and Analysis

The dimensionality reduction is reduced and its execution time decreases whenever minimum count increases shown in

Fig1.The proposed techniques are implemented by showing the execution of various algorithms runs on textual document dataset used for analyzing and the comparative analysis is shown for three different algorithms in Fig2.

Execution time decreases for frequent utility pattern agglomerative clustering compared to frequent word sequence and frequent item based on maximum occurrence. Hence WFUPAC method yields quality clusters with efficient solution with less computation time.

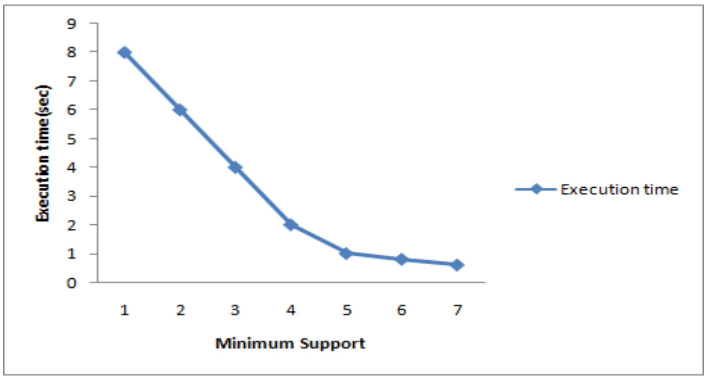

Fig1. Execution time decreases with respect to minimum support

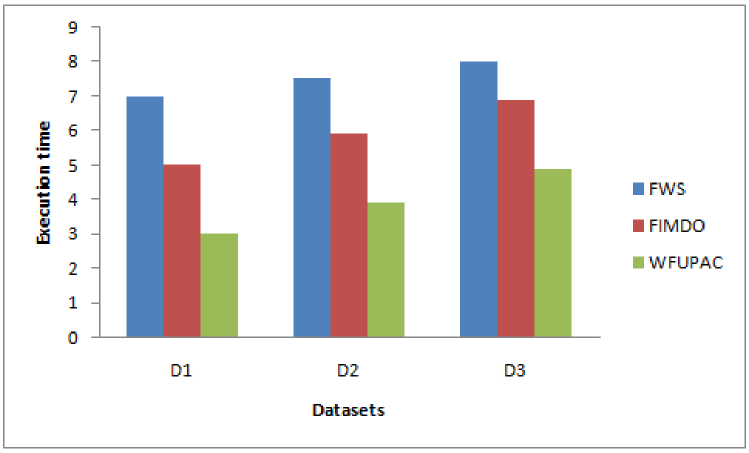

Fig2. Comparative Analysis of FWS, FIMDO, WFUPAC algorithms.

5. Conclusion:

When we compare the performance of frequent word sequence, frequent item maximum document occurrence and weighted frequent utility pattern agglomerative clustering, the weighted frequent utility pattern agglomerative clustering outperforms the other two method, thus it proves efficient clustering.

REFERENCES:

[1] Tanvir Habib Sardar, Zahid Ansari.2018. An analysis of Map Reduce efficiency in document clustering using parallel K-means algorithm. *Future Computing and Informatics Journal,* Vol. 3. pp 202 -209.

[2] K. Dhinakaran, Udhayakumar Shamugam et.al (2020), "Distributed Data Analytics for Improving Indian Economical Growth Using Recommendation System" , Jour of Adv Research in Dynamical & Control Systems, Vol. 12, 04-Special Issue, pp-134-140.

[3] Sung-Sam Hong1, Wanhee Lee, and Myung-Mook Han. The Feature Selection Method based on Genetic Algorithm for Efficient of Text Clustering and Text Classification, *Int. J. Advance Soft Compu. Appl,* Vol. 7, No. 1, March 2015, ISSN 2074-8523.

[4] Chowdam Sreedhar, Nagulapally Kasiviswanath and Pakanti Chenna Reddy. 2017.Clustering large datasets using K-means modified inter and intra clustering (KM-I2C) in Hadoop. *Journal of Big Data* .Vol. 4:27. DOI 10.1186/s40537-017-0087-2.

[5] Dr. Tamanna Siddiqui,Parvej Aalam.2015. Short Text Clustering; Challenges & Solutions: A Literature Review. *International Journal Of Mathematics And Computer Research*. Vol.3:6. pp. 1025-1031

[6] Ramzan Talib, Muhammad Kashif Hanify, Shaeela Ayeshaz, and Fakeeha Fatimax.2016. Text Mining: Techniques, Applications and Issues. *International Journal of Advanced Computer Science and Applications*, Vol. 7(11).pp.414-418.

[7] Abdolreza Hatamlou. 2013. Black hole: A new heuristic optimization approach for data clustering. *Information Sciences*. Vol. 222 (2013).pp. 175–184

[8] Shuiqiao Yang, Guangyan Huang, Borui Cai, 2019. Discovering Topic Representative Terms for Short Text Clustering, *Access IEEE*, vol. 7, pp. 92037-92047, 2019.

[9] Xia Jing. 2012. A Bayesian Network Based Intelligent Plant Classification System. *Information Science and Engineering (ISISE) 2012 International Symposium on*, pp. 263-265, 2012.

[10] EDransfield, J.FMartin, T.M Nagapo. 2004. The application of a text clustering statistical analysis to aid the interpretation of focus group interviews. *Food Quality and Preference*. Vol. 15:5 pp 477-488.

[11] CongnanLuoaYanjunLibSoon M.Chung. 2009. Text document clustering based on neighbours, *Data & Knowledge Engineering.* Vol. 68:11, pp. 1271-1288

[12] R.JananiDr.S.Vijayarani. 2019. Text document clustering using Spectral Clustering algorithm with Particle Swarm Optimization. Expert Systems with Applications. Volume 134:15 pp. 192-200. https://doi.org/10.1016/j.eswa.2019.05.030

[13] CaiyanJiaaMatthew B.CarsonbXiaoyangWanga. 2018. Concept decompositions for short text clustering by identifying word communities. *Pattern Recognition.* Vol. 76, pp. 691-703